IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁑ Impact Factor 8.102 ⁑ Peer-reviewed / Refereed journal ⁑ Vol. 12, Issue 6, June 2023

DOI: 10.17148/IJARCCE.2023.12673

Effectiveness of Wavelet Based Voice Morphing

Roshan Arun Chavan[1], Dr.Prakash Kene[2]

MCA Department,PES Modern College Of Engineering Pune,India[1,2]

Abstract: Voice morphing techniques have gained significant attention due to their applications in speech processing and multimedia systems. This study presents a wavelet-based approach for voice morphing, which enables the transformation of a source speaker's voice to resemble a target speaker while maintaining intelligibility and naturalness. The proposed method utilizes the wavelet transform to decompose the source and target speech signals into different frequency subbands. By modifying the wavelet coefficients in these subbands, the spectral and temporal characteristics of the source speech are altered to match those of the target speech. The morphed speech is then synthesized by performing an inverse wavelet transform on the modified coefficients. Objective and subjective evaluations demonstrate the effectiveness of the proposed approach in achieving accurate voice morphing while preserving the linguistic content. The waveletbased voice morphing technique presented in this study offers potential applications in speech synthesis, voice conversion, and entertainment systems.

Keywords: Voice Morphing, Speech Processing, Wavelets, Signal Decomposition, Voice Conversion

1. INTRODUCTION

Voice morphing, also known as voice transformation or voice conversion, is a technique that allows the modification of a source speaker's voice to resemble the voice of a target speaker while maintaining the linguistic content. It has found applications in various fields, including speech synthesis, multimedia systems, entertainment, and forensic audio analysis. Wavelet-based voice morphing is a

particular approach that utilizes wavelet transforms to manipulate the spectral and temporal characteristics of speech signals, providing an effective method for voice transformation.Traditional voice morphing techniques often rely on methods such as statistical modeling, Gaussian mixture models (GMMs), and hidden Markov models (HMMs) to capture the acoustic properties of speakers. While these approaches have shown promising results, they may suffer from limitations in capturing finegrained details and preserving naturalness. Wavelet-based voice morphing offers a different perspective by leveraging the benefits of wavelet analysis, which provides a time-frequency representation of signals and allows for localized frequency analysis.

Wavelets are mathematical functions that possess both time and frequency localization properties. By decomposing speech signals into different frequency subbands using wavelet transforms, it becomes possible to selectively modify specific frequency components while preserving others. This decomposition allows for more precise control over the spectral and temporal characteristics of the speech signal during the morphing process.

In wavelet-based voice morphing, the source and target speech signals are decomposed into wavelet coefficients at different scales and positions. These coefficients represent the magnitude and phase information at various time-frequency locations. By manipulating these coefficients, the spectral envelope and temporal structure of the source speech can be altered to resemble the target speech. The modified coefficients are then synthesized back into a time-domain waveform using the inverse wavelet transform.

Wavelet-based voice morphing techniques have shown promise in achieving accurate voice transformation while preserving the linguistic content of the speech. The ability to analyze speech signals in both the time and frequency domains offers more control over the morphing process, allowing for finer adjustments and improved naturalness. Furthermore, the flexibility of wavelet-based methods allows for the integration of additional processing steps, such as noise reduction, pitch modification, and voice quality enhancement, to further enhance the morphed output.

This study aims to explore and evaluate the effectiveness of wavelet-based voice morphing techniques in achieving high-quality and naturalsounding voice transformation. Objective and subjective evaluations will be conducted to assess the quality and similarity of the morphed speech compared to the target speech. The findings from this research can contribute to advancements in voice morphing technology, benefiting applications such as speech synthesis, voice conversion, multimedia systems, and entertainment

2.MAIN FEATURES OF PROPOSED METHOD

Our proposed model uses the theory of wavelets as a means of extracting the speech features followed by Radial Basis Function Neural Networks (RBFNN) for modeling the conversion.

2.1 Wavelet Decomposition Wavel

Subband decomposition is implemented using the Discrete Wavelet Transform (DWT).

Wavelets are a class of functions that possess compact support and form a basis for all finite energy signals. They are able to capture the nonstationary spectral characteristics of a signal by decomposing it over a set of atoms which are localized in both time and frequency. The DWT uses the set of dyadic scales and translates of the mother wavelet to form an orthonormal basis for signal analysis.

In wavelet decomposition of a signal, the signal is split using high-pass and low-pass filters into an approximation and a detail. The approximation is then itself split again into an approximation and a detail. This process is repeated until no further splitting is possible or until a specified level is reached. Fig. 1 shows a diagram of a wavelet decomposition tree [13]. The DWT provides a good signal processing tool as it guarantees perfect reconstruction and prevents aliasing when appropriate filter pairs are used.

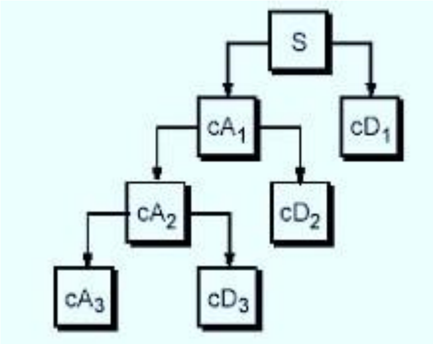

Fig. 1: Example of a wavelet decomposition tree. The original signal S is split into an approximation cA$_1$ and a detail cD$_1$. The approximation is then itself split into an approximation and a detail and so on. Decomposing a signal into k levels of decomposition therefore results in k+1 sets of coefficients at different frequency resolutions, k levels of detail and 1 level of approximation coefficients.

2.2 Radial Basis Function Neural Networks

the dimension of the input space and c the dimension of the output space. For a d-dimensional input vector x, the basic form of the mapping is

$$y_k(x) = \sum_{j=1}^{M} w_{kj}\phi_j(x) + w_{k0} \qquad (1)$$

Where □ $_j$ is the jth radial basis function (RBF) and wk 0 is the bias term which can be absorbed into the summation by including an extra RBF □$_0$ whose activation is set to 1 and □ $_j$ is usually of the form

$$\phi_j(x) = \exp\left[-\frac{1}{2}(x-\mu_j)^T\Sigma_j(x-\mu_j)\right] \qquad (2)$$

DOI: 10.17148/IJARCCE.2023.12673

where □ $_j$ is the vector determining its centre and □ j is the covariance matrix associated to □ $_j$, the width of the basis function [14].

The learning process of the RBFNN is done in two different stages. The first stage of the process is unsupervised as the input data set { x^n } alone is used to determine the parameters of the basis functions.

The selection is done by estimating a Gaussian Mixture Model (GMM) probability distribution from the input data using ExpectationMaximization [16]. The widths of the RBFs are calculated from the mean distance of each centre to its closest neighbor. The second stage of the learning process is supervised as both input and output data are required. Optimization is done by a classic least squares approach. Considering the RBFNN mapping defined in (1) (and absorbing the bias parameter into the weights) we now have

Where □$_0$ is an extra RBF with activation value fixed at 1. Writing this in matrix notation y(x)=W□ (5)

where W = (wkj) and □ = (□ $_j$). The weights are then optimized by minimization of the sum-of- squares error function

$$y_k(x) = \sum_{j=0}^{M} w_{kj}\phi_j(x) \qquad (4) \qquad\qquad E_R = \sqrt{\frac{\sum_{m=1}^{M}(y(m) - y^*(m))^2}{\sum_{m=1}^{M} y(m)^2}} \qquad (9)$$

$$E = \frac{1}{2}\sum_{n}\sum_{k}\{y_k(x^n) - t_k^n\}^2 \qquad (6)$$

where t ^{n} is the target value for output unit k whe network is presented with the input vector x^n weights are then determined by the linear equations [14]

where (T) nk = t ^{n} and (□) $_{nj}$ =□ $_j$ (xn) . This can be solved by

WT =□†T

where □† denotes the pseudo-inverse of □ . The second layer weights are then found by fast, linear matrix inversion techniques [14].

3.METHODOLOGY

Voice morphing is performed in two steps: training and transformation. The training data consist of repetitions of the same phonemes uttered by both source and target speakers. The utterances are phonetically rich (i.e. the frequency of occurrence of different phonemes is proportionate to their frequency of occurrence in the english language) and are normalized to zero mean and unit variance. The source and target training data is divided into frames of 128 samples and the data is randomly divided into training and validation sets. A 5-level wavelet decomposition is then performed to the source and target training data. The wavelet basis used is the one which gives the lowest reconstruction error given by where M is the number of points in the speech signal, m = 1,L, M is the index of each sample,y(m) is the original signal and y* (m) is the signal reconstructed from the wavelet coefficients using the Inverse Discrete Wavelet Transform (IDWT). The wavelet basis used was the Coiflet 5 for maleto-female and male-to-male speakers morphing and the Biorthogonal 6.8 for female-to-male and female-to-female speakers morphing as they gave the smallest reconstruction error.

In the transformation stage, the wavelet coefficients are calculated at different subbands. In order to reduce the number of parameters used and reduce the complexity of the RBFNN mapping, the coefficients at the two levels of highest frequencies are set to zero. At each of the remaining 4 levels, the wavelet coefficients are normalized to zero mean and unit variance (by using the

coefficients' statistics) and a mapping is learned using the RBFNN model using frames of coefficients as input vectors. The best RBFNN on each level is chosen by minimizing the error on the validation data. The complexity of the RBFNN depends on the size of the training data which varies depending on the frequency of occurrence of each phoneme in the available speech corpus. More than 20 RBF centers were rarely needed in order to optimize

DOI: 10.17148/IJARCCE.2023.12673

the network. Test data from the source speaker are then pre-processed and split into the same number of subbands as the training data and the wavelet coefficients are projected through the trained network in order to produce the morphed wavelet coefficients. The morphed coefficients are then unnormalized i.e. they are given the statistics (mean and variance) of the wavelet coefficients of the target speaker and then used to reconstruct the target speaker's speech signal. Post-processing again includes amplitude editing so that the morphed signal has the statistics of the target speaker. The process is repeated for all different phonemes of interest which are then put together in order to create the desired text. Fig. 2 shows a diagram of the proposed model.

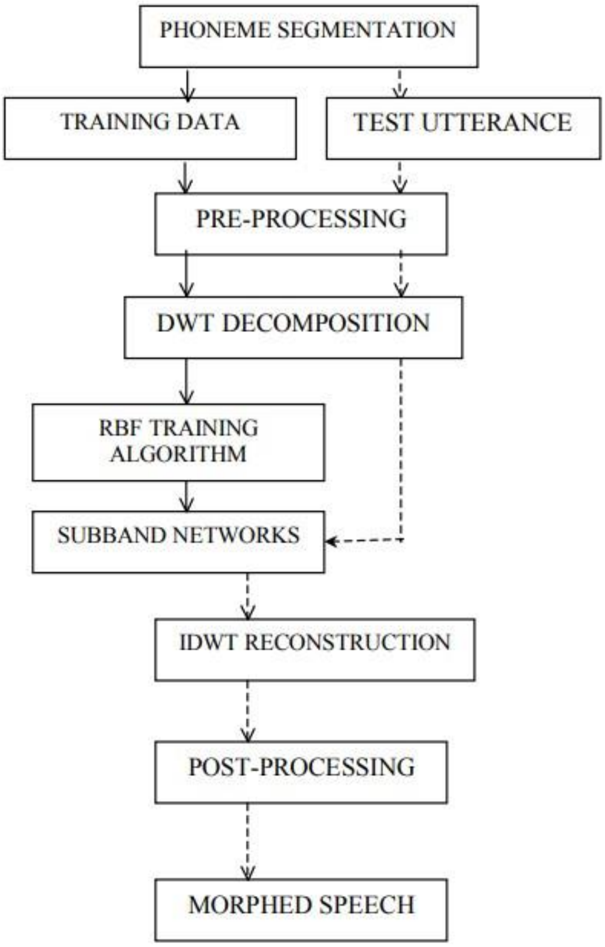

Fig 2. Proposed model

4.RESULTS AND EVALUATION

Figures 3, 4, 5 and 6 show some results of our morphing on speech from male-to-male, female-tofemale, female-to-male, and male-to-female

pairs of speakers. The sentences uttered are taken from the TIMIT database [15]. In order to evaluate the performance of our system in terms of it perceptual effects an ABX-style preference test was performed, which is common practice for voice morphing evaluation tests [3, 5]. Independent listeners were asked to judge whether an utterance X sounded closer to utterance A or B in terms of speaker identity, where X was the converted speech and A and B were the

DOI: 10.17148/IJARCCE.2023.12673

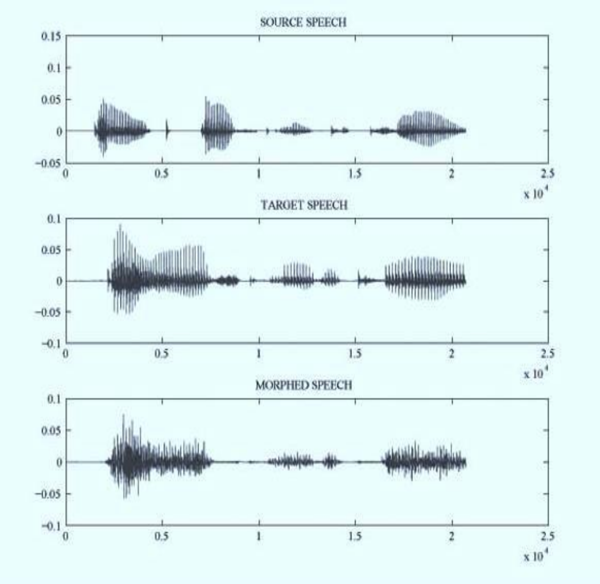

Fig 3. Source, Target and Morphed sentence waveform for a male-to-male speaker speech transformation of the utterance “Don't ask me to carry”.

source and target speech, respectively. The ABXstyle test performed is a variation of the standard ABX test since the sound X is not actually spoken by either speaker A or B, it is a new sound and the listeners need to identify which of the two sounds it resembles. Also, utterances A and B were presented to the listeners in random order. In total, 16 utterances were tested which consisted of 4 male-tomale, 4 female-to-female, 4 female-to-male and 4 male-to-female source-target combinations. All utterances were taken from the TIMIT database. 11 independent listeners took part in the testing. Each listener was presented with the 16 different triads of sounds (source, target and converted speech, the first two in random order) and had only one chance of deciding whether sound X sounds like A or

B. It is assumed that there is no correlation between the decisions made by the same person and that all 176 resulting decisions are independent. Since the probability of a listener recognizing the morphed

speech as the target speaker is 0.5, the results were verified statistically by testing the null hypothesis that

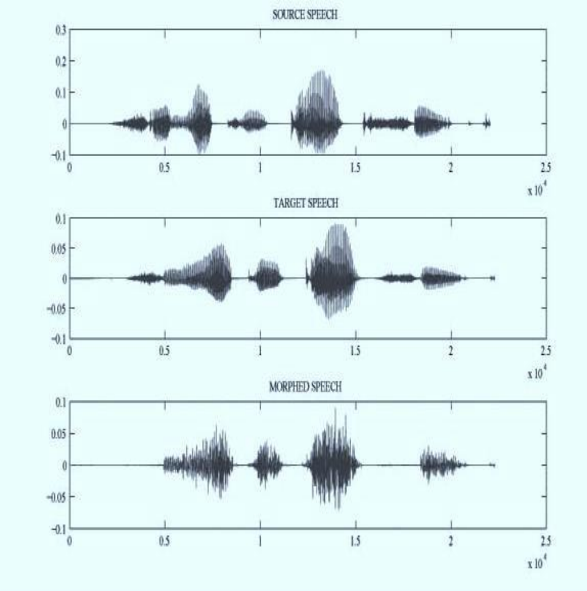

Fig 4. Source, Target and Morphed sentence waveform for a female-to-female speaker speech transformation of the utterance “Don't ask me to carry”the probability of recognizing the target speaker is 0.5 versus the alternative hypothesis that the probability is greater than 0.5. The measure of interest is the p-value associated with the test i.e. probability that

DOI: 10.17148/IJARCCE.2023.12673

the observed results would be obtained if the null hypothesis was true i.e. if the probability of recognizing the target speaker was 0.5. Table 1 gives the percentage of the converted utterances that were labeled as closer to the target speaker as well as the p-values.

Source-Target	% success	p-value
Male-to-Male	79.5	0.0001
Female-to Female	77.3	0.0003
Male-to-Female	86.3	0.00004
Female-to-Male	88.6	0.00001

Table 1. Percentage of "successful" labeling and associated p-values.

The p-values obtained are considered statistically insignificant, it is therefore evident that the null hypothesis is rejected and the alternative hypothesis is valid i.e. the converted speech is successfully recognized as the target speaker.

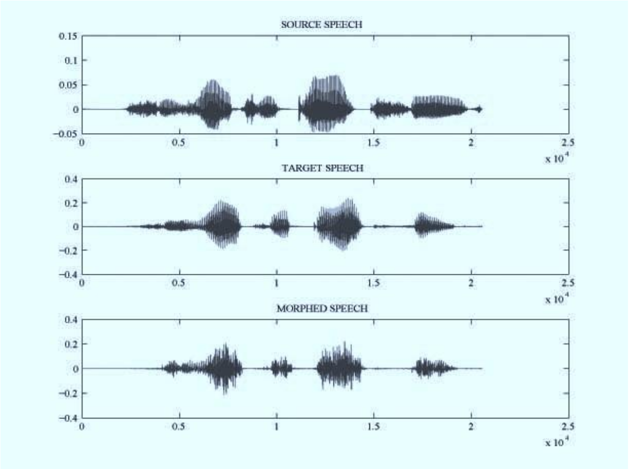

Fig 5: Source, Target and Morphed sentence waveform for a female-to-male speaker speech transformation of the utterance "She had your dark suit".

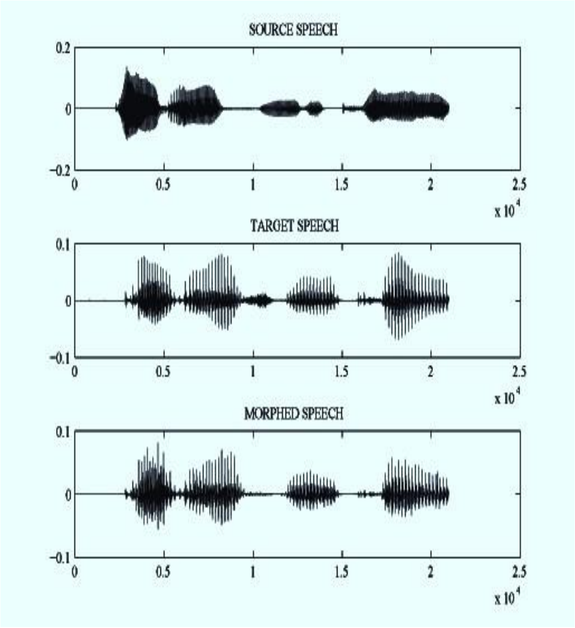

Fig 6. Source, Target and Morphed sentence waveform for a male-to-female speaker speech transformation of the utterance "She had your dark suit".

DOI: 10.17148/IJARCCE.2023.12673

5.ADVANTAGES AND DISADVANTAGES

Advantages of wavelet-based voice morphing:

Time-frequency localization:

Wavelets provide excellent time-frequency localization properties, allowing precise control over the transformation of different frequency components of the voice signal. This enables more accurate manipulation of specific features such as pitch, timbre, and formants.

Multiresolution analysis:

Wavelets offer a multiresolution analysis, which means they can represent the voice signal at different scales or resolutions. This allows for selective modification of specific scales, providing greater flexibility in voice morphing. Different characteristics of the voice can be altered independently, resulting in a more naturalsounding morphed voice.

Seamless transitions:

Wavelets can ensure smooth transitions between different parts of the voice signal, minimizing artifacts or discontinuities that can occur in other morphing techniques. This helps in achieving a more natural and coherent transformation between the original and morphed voices.

Reduced computational complexity:

Compared to some other signal processing techniques, wavelet-based voice morphing can offer lower computational complexity. This makes it more efficient for real-time or near real-time applications where low latency is desired.

Disadvantages of wavelet-based voice morphing:

Complexity of implementation:

Wavelet-based voice morphing requires expertise in signal processing and a thorough understanding of wavelet theory. Implementing the technique can be complex and may require specialized software or libraries, limiting its accessibility to users without the necessary technical knowledge.

Limited control over specific voice features:

While wavelets provide good control over timefrequency characteristics, they may not offer as fine-grained control over certain voice features compared to other methods. For example, morphing specific phonemes or accent characteristics might be more challenging with wavelet- based techniques.

Potential for artifacts:

Although wavelets can minimize artifacts, there is still a possibility of introducing some distortions or artifacts in the morphed voice. These artifacts can manifest as unwanted noise, phase inconsistencies, or undesired changes in the spectral characteristics of the voice.

Complexity of parameter selection:

Adjusting the parameters of wavelet-based voice morphing algorithms can be non-intuitive and require experimentation. Finding the optimal set of parameters for a desired voice transformation can be time-consuming and may involve trial and error.

6.APPLICATION OF WAVELET BASED VOICE MORPHING

Voice transformation in entertainment industry:

Wavelet-based voice morphing techniques can be used in the entertainment industry for audio postproduction and voice modification in movies, television shows, and video games. It enables altering voices to create special effects, simulate different characters, or modify vocal characteristics to match specific requirements.

DOI: 10.17148/IJARCCE.2023.12673

Speech synthesis and voice conversion:

Wavelet-based voicemorphing can be utilized in speech synthesis systems to converttext into speech with modified voice characteristics. It allowsfor adjusting the pitch, timbre, and other vocal attributes togenerate synthetic speech that sounds natural and personalized.Voice conversion applications can use wavelet-based morphingto transform one person's voice into another, which findsapplications in voice dubbing, language learning, and voicebased personalization.

Vocal training and therapy:

Wavelet-based voice morphing assist in vocal training and therapy sessions. It enables voice professionals, singers, or individuals undergoing speech therapy to analyze and modify their vocal performance. By selectively modifying specific features like pitch, formants, or vibrato, voice morphing can aid in improving vocal skills,correcting pronunciation, or providing a therapeutic tool for individuals with speech disorders.

Forensic analysis:

Wavelet-based voice morphing play a role in forensic analysis, particularly in voice authentication and speaker verification. By morphing a suspect's voice to match a known reference voice, investigators can assess the likelihood of voice tampering or impersonation in criminal cases. It provides a means to analyze voice evidence and evaluate the reliability of audio recordings.

Human-computer interaction and virtual assistants:

Wavelet-based voice morphing techniques can enhance the interaction between humans and virtual assistants, chatbots, or voice-controlled systems. By morphing the voice output of these systems, they can be customized to have different voice characteristics or match the user's preferences. This personalization can lead to a more engaging and natural user experience.

Audio watermarking and copyright protection:

Wavelet-based voice morphing can be used for audio watermarking, embedding imperceptible and robust information within voice signals. This information can serve as a digital watermark for copyright protection, ownership verification, or tracking purposes. Wavelet-based morphing techniques can aid in embedding and extracting watermarks while maintaining the integrity of the audio signal.

7. CONCLUSION

A voice morphing system was presented which extracts the voice characteristics by means of wavelet decomposition and then uses the theory of RBFNN for morphing at the different levels of decomposition. The experimental results show that the conversion is successful in terms of producing speech that can be recognized as the target speaker although the speech signals sounded muffled. Furthermore, it was observed that most distortion occurred at the unvoiced parts of the signals. The muffled effect as well as the distortion could be due to the removal of some of the highest frequency components of the speech signals therefore methods of including all frequency levels to the morphing will be a subject of future work.

REFERENCES

[1]. Drioli C., Radial basis function networks for conversion of sound speech spectra, EURASIP Journal on Applied Signal Processing, Vol. 2001, No.1, 2001, pp. 36-40.

[2]. Valbret H. , Voice transformation using PSOLA technique , Speech Communication, Vol .11, No 23, 1992, pp. 175-187.

[3]. Arslan L., Speaker transformation algorithm using segmental codebooks, Speech Communication, No. 28, 1999, pp. 211-226.

[4]. Arslan L. and Talkin D, Voice Conversion by Codebook Mapping of Line Spectral Frequencies and Excitation Spectrum, Proceedings of Eurospeech, 1997, pp. 1347-1350.

[5]. Stylianou Y., Cappe O. and Moulines E., Statistical Methods for Voice Quality Transformation, Proceedings of Eurospeech, 1995, pp. 447-450.